Adquisición de conocimiento léxico: Acquilex.

A. Ageno (1) I. Castellón (1) M.A. Martí (2) F. Ribas (1) G. Rigau (1)
H. Rodriguez (1) M. Taulé (2) F. Verdejo (3)

(1) Universitat Politècnica de Catalunya. Departament de LSI. Barcelona.
(2) Universitat de Barcelona. Departament de Filologia Romànica. Barcelona
(3) UNED Departamento de Ingeniería Electrica, Electrónica y de Control. Madrid

1. Introducción.

El objetivo básico del proyecto **ACQUILEX** es el desarrollo de técnicas y métodos que permitan la utilización de diccionarios en soporte magnético (M.R.D, Machine Readable Dictionaries) en la construcción de componentes léxicas para sistemas de procesamiento del lenguaje natural (P.L.N).

Los diccionarios automatizados constituyen una fuente de adquisición de Conocimiento Léxico y conceptual que, potencialmente, permite abordar algunos aspectos especialmente costosos de la construcción de una base de conocimiento para un sistema de P.L.N. en forma rápida y competitiva. Se trata de un campo relativamente poco explorado del área de la Adquisición del Conocimiento, debido a la dificultad que supone el tratamiento complejo de grandes volúmenes de información y a las limitaciones de las teorías lingüísticas que abordan el tema del léxico.

A largo plazo el objetivo del proyecto es la construcción de una Base de Conocimientos léxicos multilingüe con las siguientes caraterísticas:

- Contendrá Información Léxica general e independiente del dominio.

(*) **ACQUILEX** *es un proyecto integrado, en el que participan el Instituto de Lingüística Computacional de Pisa y las universidades de Amsterdam, Cambridge, Dublin y Politécnica de Catalunya . El proyecto esta financiado por la C.E.E. a través del programa ESPRIT (Acción BRA 3030).*

• La Representación del Conocimiento favorecerá al máximo su reutilización.

• Se utilizarán exclusivamente fuentes léxicas ya existentes.

• Los procesos de extracción de la información léxica y de utilización de la misma por los diferentes sistemas de tratamiento del Lenguaje Natural serán distintos e independientes.

• Se utilizará un formato estandar de intercambio de fuentes léxicas.

• Se definirá una estructura conceptual común, ligada a los significados individuales de las palabras en las diferentes lenguas cubiertas y capaz de soportar un procesamiento del lenguaje basado en el Conocimiento.

• Se incluirá un vocabulario general con información fonológica, morfológica, sintáctica y semántico-pragmática para las diversas lenguas que forman parte del proyecto.

Los objetivos primeros del proyecto se centran en el desarrollo de un prototipo de Base de Datos Léxica (L.D.B) y de Base de Conocimientos Léxica (L.K.B.) multilingües para un subconjunto manejable, pero significativo, del vocabulario y en el desarrollo de técnicas para la extracción semiautomática de información léxica del diccionario.

Las etapas principales en el desarrollo del proyecto, buena parte de las cuales ya han sido cubiertas, son las siguientes:

a. Elaboración de un modelo computacional del diccionario que describa su contenido de forma que tengan expresión en él todas las características diferenciales de los diferentes diccionarios individuales.

b. Descripción de los diccionarios individuales en términos del modelo.

c. Definición de una L.D.B. Desarrollo de software de gestión de dicha L.D.B.

d. Carga de la información de los diccionarios individuales en la L.D.B.

e. Derivación de una estructura conceptual común. Relaciones entre esta estructura y las definiciones individuales de cada diccionario.

f. Desarrollo de técnicas para la extracción de información conceptual a partir de la información léxica contenida en la L.D.B.

g. Carga en la L.K.B. de un subconjunto significativo de la información léxica de los diversos diccionarios individuales.

h. Chequeo y evaluación del sistema a través de la actuación de un Sistema de P.L.N. cuya componente léxica se haya extraído de la L.K.B. El Sistema chequeará las dos funciones de comprensión y generación.

2. Carga del MRD en la LDB.

El diccionario fuente del que se parte es un texto continuado que contiene dos tipos de información: el texto del diccionario y códigos tipográficos. El primer paso que se debe realizar para acceder a la información del diccionario es la construcción de una gramática que lo analice y segmente por entradas y cada entrada por diferentes campos que contiene.

Ejemplo de la entrada 'heraldo' del diccionario Vox en cinta :

EP[j2]heraldo [k1](fráncico [k2]heriald, [k1]miembro del
ejército; a través del fr. [k2]héraut[k1]) [k2]m. [k1]Oficial que en la
Edad Media tenía a su cargo transmitir mensajes, ordenar las fiestas de
caballería, llevar los registros de la nobleza, etc.[k2] 2 [k1]fig.
Mensajero, adalid.[EP[j3] [j6]Sin.[j7] [k2]1
[k3]Faraute.

El resultado de la gramática es una estructura parentizada donde la información aparece etiquetada.

Ejemplo de la entrada 'heraldo' del diccionario Vox en formato parentizado:

((heraldo)
(ETIM fráncico heriald , miembro del ejército; a través del fr. héraut)
(Sense 1)
(CA m.)
(DEF Oficial que en la Edad Media tenía a su cargo transmitir mensajes, ordenar las fiestas de caballería, llevar los registros de la nobleza, etc.)
(Sense 2)
(CA m.)
(SEM fig.)
(DEF Mensajero, adalid.)
(RELA 1)
(TIPOR Sin.)
(TXR 1 Faraute.)
)

El siguiente paso es el volcado de esta estructura parentizada a la LDB. Cada entrada debe indiciarse para poder acceder de forma rápida a la información que contiene.

Ejemplo de la entrada *heraldo (1)* del diccionario Vox en formato LDB :

```
[[SIN
  [CA m.]]
 [SEM
  [DEF oficial que edad media tenía cargo transmitir mensajes
    ordenar fiestas]]
 [FORMA
  [ETIM fráncico heriald , miembro ejército; a través del fr.
    héraut]
  [FORM
    [TIPOF ]
    [TXF ]]]
 [RELA
  [TIPOR Sin.]
  [TXR Faraute. 1]]]
```

3. Construcción de taxonomías.

3.1. Nomenclatura.

Taxonomía

La taxonomía es un paso intermedio entre la LDB y la LKB.
Es una clasificación jerárquica de sentidos del diccionario que están conectados entre ellos mediante la relación ES-UN.

Raíz de taxomía ('top')

El sentido que inicia la taxonomía, a partir del cual empieza la búsqueda de hipónimos.

Terminal de taxonomía ('bottom')

Sentido que ya no apareceen ninguna definición del diccionario como término génerico.

Término genérico o hiperónimo

Es el elemento nuclear de la definición. Indica a qué clase pertenece el objeto definido.

Hipónimo

Es el sentido definido mediante un término genérico determinado.

Fusión

Operación de simplificación de sentidos de una entrada, se realiza cuando dos acepciones describen un mismo concepto o conceptos muy cercanos.

Desambiguación

Operación que consiste en determinar a que acepción de una entrada se refiere una ocurrencia determinada de esa entrada.

Para su realización de modo semiautomático se requiere la determinación de una serie de

heurísticos.

Heurísticos

Implementan estrategias para la toma de decisiones allí donde no está definida ninguna solución algorítmica. Cada heurístico es un procedimiento que asigna una puntuación a cada una de las diferentes opciones que se le plantean y selecciona aquella de valoración más alta.

3.2. Herramientas utilizadas.

Adaptación y optimización del analizador morfológico **SegWord** (Sanfilippo).

Adaptación del analizador sintáctico-semántico **FPar** (Ashawi).

Desarrollo de **SEISD** (Sistema para la Extracción de Información Semántica de Diccionarios) para la extracción de:

-relaciones semánticas entre los sentidos del diccionario VOX : taxonómicas y meronímicas,

- extracción de las 'diferentiae' (modificadores del término genérico).

3.3. Proceso de construcción de la taxonomía.

1.- Determinación del 'top' y fusión de sus sentidos

2.- Búsquedas en la LDB para obtener todos aquellos sentidos que :

- tienen la misma categoría que el ´top´ seleccionado.

```
[[SIN
  [CA (OR s.pl. s. m.pl. m. f.pl. f. adj.-s. adj.-m. adj.-f.)]]]
```

- incluyen en su definición la palabra elegida como 'top'

[[SEM
[DEF substancia]]
[SIN
[CA (OR s.pl. s. m.pl. m. f.pl. f. adj.-s. adj.-m. adj.-f.)]]]

3.- Análisis morfológico y sintáctico de las definiciones obtenidas para obtener todos aquellos sentidos que tienen como genérico el 'top' determinado.

entrada

abono [de abonar I] ** I
acepción:1 ** m. ** Substancia mineral u orgánica que se
añade a la tierra para fertilizarla.

analisis morfológico

((SUBSTANCIA V N) (MINERAL ADJ N) (U CONJ N) (ORGáNICA ADJ) (QUE CONJ) (SE PRON) (AñADE V) (A P) (LA DET) (TIERRA N) (PARA P) (FERTILIZARLA V))

analisis sintáctico

(((CLASS SUBSTANCIA) (RELATED-TO MINERAL) (PROPERTIES ORGáNICA) (GOAL FERTILIZARLA)))

4.- Desambiguación de sentidos: asociar cada hipónimo al sentido del hiperónimo correspondiente.

5.- Repetición del proceso (puntos 2, 3 y 4) para los hipónimos obtenidos y así sucesivamente hasta hallar los terminales de la taxonomía.

4. Proyección a la LKB.

Una vez extraída la información semántica de las definiciones del diccionario, la siguiente etapa consiste en proyectar esta información a la LKB (Base de Conocimiento Léxico). Para realizar esta operación hemos desarrollado un entorno de conversión que transforma cada

nodo de la taxonomía y su información semántica asociada a su correspondiente entrada léxica de la LKB.

El lenguaje de representación de Conocimiento que empleamos está basado en la unificación tipada e incorpora herencia por defecto (Carpenter 1990).

Actualmente, estamos diseñando la estructura de 'qualias' de los nombres, donde quedarán reflejadas las propiedades de los conceptos.

4. Referencias.

[Ageno et al. 92] Ageno A., Castellón I., Martí M. A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo M. F. "SEISD: An environment for extraction of Semantic Information from on-line dictionaries.". *Proceedings 3rd Applied Natural Language Processing. Trento. Italy.*

[Alshawi 89] Alshawi H. "Analysing the dictionary definitions". In Boguraev B., Briscoe T. (eds) *Computational Lexicography for NLP* , chapter 7. Longman, London.

[Carpenter 90] "Typed feature structures: Inheritance, (In)equacy and Extensionaly". *Proceedings of the First International Workshop on Inheritance in NLP*, Tilburg, The Netherlans, pp. 9-18.

[Copestake 90] Copestake A. "A System for building disambiguated taxonomies". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO.012

[Sanfilippo 90] Sanfilippo A. "A morphological Analyser for English & Italian". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO. 004